ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

基于共现关系的关键词层次结构构建研究

熊回香 陈子薇 叶佳鑫
华中师范大学信息管理学院 武汉 430079

摘要：[目的/意义]关键词作为应用最为广泛的文献知识单元，对于其语义关系的深入挖掘可为知识关联、资源推荐等工作提供底层支持。[方法/过程]基于关键词直接共现与间接共现关系对关键词之间的相关性进行挖掘，在此基础上对关键词的分布情况进行分析并结合关键词概念范围大小构建关键词间的层次结构。[结果/结论]以“知识图谱”为根节点，演示关键词层次结构构建步骤，研究表明，该方法具有一定的可行性和有效性，能够较好地构建关键词层次结构。

关键词：科技文献关键词　关键词层次结构　关键词特征

分类号：G25；TP391.1

引用格式：熊回香，陈子薇，叶佳鑫. 基于共现关系的关键词层次结构构建研究 [J/OL]. 知识管理论坛，2022, 7(4): 443-451[引用日期]. http://www.kmf.ac.cn/p/306/.

1 引言

科技文献主要包括题名、关键词、摘要、全文等重要内容，其中关键词是最常用的表示科技文献内容特征的知识单元，相较于题名来说关键词能表示文本内容特征的不同侧面，与摘要相比关键词表示的知识则更为浓缩，相较于全文来说关键词则具有利用便捷、高效的特点 [1-3]。由此关键词成为目前应用范围最广、最受关注的科技文献知识单元。

对于关键词的开发利用，主要是在关键词间相关性挖掘的基础上，借助关键词来表征文本、资源或者使用关键词的用户的特征，进而通过关键词之间的关联来建立文本间、资源间以及用户间的联系，实现知识关联、资源推荐等工作。早期，关键词间相关性的挖掘主要依赖于对词典资源的利用，但因词典存在更新速度慢、覆盖面有限等问题，关键词间相关性的挖掘逐步转向于从大规模的语料库中学习并构建关键词特征，通常采用向量特征来计算关键词间相似度 [4]。然而，关键词间存在着同义、上下义、反义、同形异义等多种关系，目前的

基金项目：本文系国家社会科学基金年度项目“融合知识图谱和深度学习的在线学术资源挖掘与推荐研究”（项目编号：19BTQ005）研究成果之一。
作者简介：熊回香，教授，博士生导师，博士；陈子薇，硕士研究生，通信作者，E-mail: vv@mails.ccnu.edu.cn；叶佳鑫，博士研究生。
收稿日期：2022-03-15　发表日期：2022-08-22　本文责任编辑：刘远颖

研究常将这些复杂的关系以单一的相似度数值来度量，例如基于关键词共现关系的词间相似度挖掘，这种方法并未对不同关系进行区分，缺少对关键词语义信息的深入挖掘，也导致在效果上存在一定不足 [5-6]。笔者从科技文献价值开发的角度出发，在关键词共现分析的基础上结合对词本身分布特征的分析，建立能反映关键词间研究范围上下位关系的关键词层次结构，以更好地对关键词进行挖掘利用，推动相关研究进展。

2 相关研究

2.1 词语相关性挖掘

（1）基于词典的挖掘。基于词典对词语进行相关性挖掘主要是依据构建词典时的分类规则来挖掘词语之间的语义联系。WordNet 是最常见的用于挖掘英文词语间相关性的语义词典，通过 WordNet 可有效挖掘词语之间概念关系，并用于文档或图像等资源间相似度的计算 [7]；同义词词林是一本包含词语间同义关系的语义词典，其按照词语概念的递进分为 5 层树状结构，基于词林的树状结构能对词语关系进行挖掘 [8]；HowNet 也是常见的用于挖掘中文词语相关性的词典，区别于应用词林时基于词典结构，在利用 HowNet 进行词语间相关性挖掘时主要是依据描述词语概念的义原 [9]；此外，同时借助多种词典进行词语相关性挖掘，相较于借助单一词典能在一定程度上扩大可计算词语的范围并提升相关性挖掘的准确性 [10]。

（2）基于大规模语料的挖掘。相较于基于词典的方法，基于大规模语料的挖掘效果主要取决于文本特征的学习与表示方法，且其囊括的词语范围远高于基于词典的范围。目前，文本特征的学习与表示，主要是将文本特征经过训练转换为词向量，常见的主要有基于 CBOW、Skip-gram 等算法训练得到 Word2vec 词向量 [11]，以及目前较为流行的基于 CNN、LSTM 与 BERT 等模型训练得到词向量或挖掘文本特征 [12-13]。田星等将 Jaccard 与 Word2vec 相结合，在训练得到 Word2vec 词向量后，将词向量融入到 Jaccard 方法中，进行短文本间相关性挖掘，有效提升了挖掘效果 [14]；E. L. Pontes 等使用 CNN 解析单词的局部上下文，使用 LSTM 分析句子的全局上下文，对文本信息进行有效保留以提高相关性挖掘效果 [15]；M. M. Sanjeev 等借助 BERT 实现词、句子间语义相关性的挖掘，并将其应用于邮件查找工作中 [16]。

在词语相关性挖掘方法中，基于词典的方法对词语之间关系的挖掘较为全面，词语相关性挖掘的效果通常较好，但存在词典更新困难、计算范围有限的问题；而基于大规模语料的方法，虽然能显著提升计算范围，并能实现对词语关系的自动挖掘，但这类方法通常对语料的质量要求较高，且对部分词语如低频词、凸现词的挖掘效果较差 [17]。

2.2 词语层次关系挖掘

词语层次关系挖掘主要是对词语之间的上下位关系进行挖掘与呈现，即在词语相关性挖掘的基础上进一步得出词语之间的上下级关系并建立相应的词语结构，目前常见的挖掘对象主要为社交网络上的标签类词语以及学术文献中的关键词类词语。G. Tibély 等以蛋白质功能标签与电影标签为对象，基于复杂网络理论，通过网络加权与共现关系从网络中提取出了标签层次关系 [18]；S. Li 等基于学术关键词的共现关系以及词组中词的组合顺序建立了关键词层次结构 [19]；熊回香等依据图书标签的概念范围及共现关系进行了标签层次关系建立 [20-21]。

在词语层次关系构建研究中，以往的研究多以共现关系为基础进行词语之间层次关系的挖掘，但在挖掘时仅考虑了词语是否共现，没有对词语的语义类型与功能进行区分，因而难以说明层次关系是按照何种规则进行层次递进，也导致了构建的层次关系在应用上存在一定的局限。

3 研究框架与关键步骤

3.1 研究框架

为了更好地挖掘词语相关性，笔者借助词

ChinaXiv合作期刊
知识管理论坛，2022（4）:443-451
DOI: 10.13266/j.issn.2095-5472.2022.037

典的思想，对词语间共现情况进行深入挖掘来半自动地构建能反映词语间上下位关系的词语层次结构，并将建立好的层次结构与基于语料的方法结合，以拓展相关性挖掘的范围，提高挖掘结果质量。因学术关键词具有规范、精炼、语义明确等特点，笔者选择研究的词语为学术关键词，其按语义类型及功能的不同可以分为研究方法类、研究主题类、研究范围类等不同类型的关键词 [22]。其中，研究方法类关键词反映的是科技文献所用研究方法，通过挖掘不同科技文献在研究方法上存在的异同之处可以较好地挖掘文献之间的联系，并且通过研究方法之间的关联可以有效扩充研究方法的适用范围。因此，笔者在挖掘词语相关性时以研究方法类学术关键词作为主要研究对象，通过挖掘研究方法类关键词与其他类型关键词间的共现关系来构建研究方法类关键词层次结构，若某一研究方法类关键词与多种研究主题或研究范围类关键词具有共现关系，则可推断该方法适用于多种主题，具有较为广泛的应用范围，以此为基础构建研究方法类关键词的层次结构，则可按关联的主题与研究范围大小进行关键词层次递进，使构建的层次结构具有更好的应用价值。按此思路构建的研究框架共分为数据收集与预处理、关键词相似度计算、建立关键词层次结构 3 个步骤，如图 1 所示：

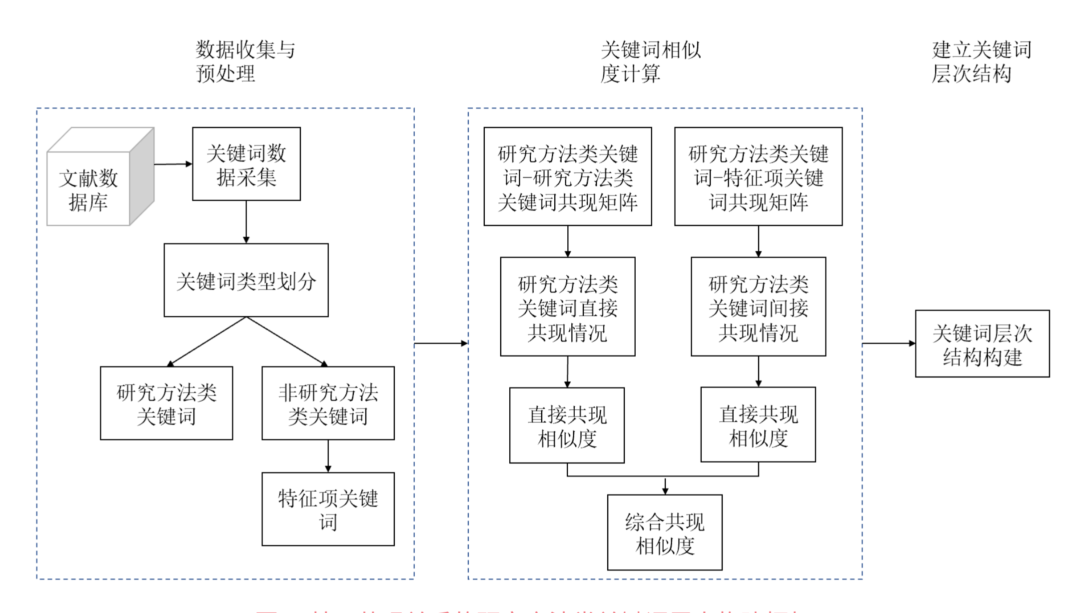


图 1 基于共现关系的研究方法类关键词层次构建框架

3.2 关键步骤

3.2.1 数据收集与预处理

从文献数据库中采集相关科技文献关键词数据，对采集到的关键词数据进行筛选与统计工作之后，按照参考文献 [3] 与参考文献 [22] 所述标准将关键词划分为研究方法类关键词与非研究方法类关键词。然后，对于非研究方法类关键词，按照词频排序，选择词频数较高的部分研究主题类与研究范围类关键词作为特征项关键词，用以在后续研究中描述研究方法类关键词的特征。

3.2.2 关键词相似度计算

基于关键词共现矩阵计算关键词间相似度。关键词间的共现可分为直接共现情况与间接共现情况两种，在本文中直接共现情况是指两个研究方法类关键词出现在同一科技文献中，即在该科技文献中两个研究方法类关键词被用于同一研究；间接共现情况则是指两个研究方法



类关键词被用于同一个研究主题或者研究范围中。笔者构建研究方法类关键词之间的共现矩阵用以反映研究方法类关键词间的直接共现情况，构建研究方法类关键词与特征项关键词之间的共现矩阵用以反映研究方法类关键词间的间接共现情况，并在共现矩阵的基础上利用余弦相似度算法计算研究方法类关键词之间的向量余弦距离，得到研究方法类关键词之间的直接共现相似度与间接共现相似度，此外，考虑到本文研究重点为关键词层次结构的构建，故直接对两种相似度进行加权整合得到研究方法类关键词综合共现相似度。

3.2.3 建立关键词层次结构

研究方法类关键词层次结构的建立主要可以分为概念范围度量、确立根节点、选定概念范围阈值、确立子节点与层级递进 5 个步骤。

（1）概念范围度量。研究方法类关键词概念范围是通过其与特征项关键词之间的共现关系度量，在本文中特征项关键词是反映文献研究主题、研究对象等特征的词，若相关的特征项关键词越多，则表明研究方法类关键词可适用于更多的研究主题或对象，具有较大的概念范围。

（2）确立根节点。根节点概念范围越大，则与其相关的关键词层次结构也能具有更大的适用范围，因此在度量研究方法类关键词概念范围之后选择概念范围较大的关键词作为层次结构的根节点。

（3）制定概念范围阈值。为使概念范围接近的关键词尽可能位于同一层级，其处于上下层级的关键词间概念范围存在一定差异，使得概念范围随着层级递进呈现逐层递减，需要控制不同层级中关键词的概念范围。故在建立层次结构时，应在对关键词概念范围进行度量的基础上，分析关键词概念范围的分布，并以此制定每个层级的概念范围阈值。

（4）确立子节点。确立根节点并制定概念范围阈值之后，按照根节点关键词与其他关键词之间的关系确立可加入层次结构的子节点关键词。首先，加入层次结构的子节点应与根节点具有一定的相关性，在本文中即子节点与根节点之间的综合共现相似度应达到一定值；其次，子节点应与某一父节点具有一定的相关性，在本文中即子节点与父节点之间的直接共现相似度或间接共现相似度应达到一定值；最后，子节点的概念范围应达到对应层级的概念范围阈值。

（5）层级递进。确立根节点后，为根节点加入子节点作为层次结构的第二层级；随后，将加入的子节点作为第三层级关键词的父节点并为其加入对应的子节点，并通过衡量关键词之间的相似度以及关键词概念范围是否达到对应的阈值来向层次结构中逐渐加入新的节点，每个关键词仅能加入层次结构 1 次，若子节点同时与多个父节点间的相似度达到阈值，则将其与相似度最大的父节点建立层次关系，且子节点概念范围应低于父节点。

4 实证研究与结果分析

4.1 数据收集及预处理

考虑到学科内的研究方法在较短年限内不会发生太大变化以及期刊在选题上具有一定的连贯性，因此选取《图书情报工作》《情报理论与实践》《情报杂志》《情报科学》《情报学报》《数据分析与知识发现》6 种与研究方法较为相关的期刊[23]作为数据来源期刊，选择“实验法”“实证研究”“统计分析”等 55 个使用频次最高的研究方法类关键词[23]作为研究对象。

在中国知网上构造检索表达式，设置源期刊为《图书情报工作》等 6 种情报学核心期刊，包含关键词为“实验法”或“实证研究”等 55 个关键词，发表时间为 2016 年 7 月至 2021 年 6 月，共检索到相关文献 1 489 篇，如表 1 所示（仅展示关键词与题名信息）：

知识管理论坛，2022（4）:443-451
DOI: 10.13266/j.issn.2095-5472.2022.037

表 1 科技文献数据

序号	题名	关键词
1	融合PageRank与评论情感倾向的在线健康社区用户影响力研究	PageRank;情感倾向;在线健康社区;用户影响力
2	基于内容的科技文献大数据挖掘与应用	科技文献;领域知识图谱;碎片化;开放数据;数据挖掘;大数据
3	基于模块化理论的复杂产品知识管理模型研究	企业;知识管理;复杂产品;知识模块化;案例分析
4	双边网络冲突结果与相对网络能力强弱相关性研究——基于双边网络事件争端数据库和贝尔弗国家网络能力指数2020	通用信息系统;专用控制系统;非对称进攻优势;相关性分析
5	学者学术生命视角下评价指标的实证研究	PageRank;情感倾向;在线健康社区;用户影响力
……	……	……
1 489	国际图书情报领域研究的前沿主题及其演化趋势分析	图书情报领域;研究主题;知识图谱;前沿趋势;CiteSpace

收集数据后，对关键词进行标准化与筛选等工作。首先将关键词进行同义词标准化，例如将“K-means”“k-means 聚类”与“K-means 算法”统一表述为“K-means”；随后进行词频统计，由于词频数较低的词构建层次关系时会使得关键词相关性难以挖掘，因此去除词频数在 5 以下的研究方法类关键词；最后根据词频选择作为特征项的关键词，选取词频数达到 9 的非研究方法类关键词作为特征项。对关键词进行筛选与选择后共有 40 个研究方法类关键词、48 个特征项关键词，分别如表 2 和表 3 所示：

表 2 研究方法类关键词词频

序号	关键词	词频
1	知识图谱	137
2	可视化	75
3	社会网络	72
4	本体	71
5	专利分析	68
……	……	……
40	模糊综合评价	6

对关键词进行词频统计与筛选后，应用 Co-Occurrence6.7(COOC6.7)[24] 构建关键词共现矩阵。根据表 1 构建表 2 中研究方法类关键词之间的共现矩阵；基于表 1 并依据表 2 中研究方法类关键词与表 3 中的特征项关键词间的共现关系，构建研究方法类关键词与特征项关键词共现矩阵。

表 3 特征项关键词词频

序号	关键词	词频
1	文献计量	55
2	网络舆情	47
3	大数据	41
4	影响因素	32
5	高校图书馆	26
……	……	……
48	信息行为	9

4.2 直接共现相似度

基于研究方法类关键词的共现矩阵，以余弦相似度算法计算研究方法类关键词向量之间的余弦距离，通过余弦距离来度量关键词之间的直接共现相似度，结果如表 4 所示：

表 4 研究方法类关键词直接共现相似度

	知识图谱	可视化	社会网络	本体	专利分析	……	模糊综合评价
知识图谱	1.000	0.169	0.449	0.493	0.244	……	0.000
可视化	0.169	1.000	0.396	0.555	0.577	……	0.000
社会网络	0.449	0.396	1.000	0.357	0.585	……	0.000
本体	0.493	0.555	0.357	1.000	0.520	……	0.000
专利分析	0.244	0.577	0.585	0.520	1.000	……	0.000
……	……	……	……	……	……	……	……
模糊综合评价	0.000	0.000	0.000	0.000	0.000	……	1.000

4.3 间接共现相似度

根据研究方法类关键词与特征词的共现矩阵，以余弦相似度算法计算研究方法类关键词向量之间的余弦距离，通过余弦距离来度量关键词之间的间接共现相似度，结果如表 5 所示：

表 5 研究方法类关键词间接共现相似度

	知识图谱	可视化	社会网络	本体	专利分析	……	模糊综合评价
知识图谱	1.000	0.736	0.494	0.261	0.203	……	0.189
可视化	0.736	1.000	0.667	0.306	0.257	……	0.214
社会网络	0.494	0.667	1.000	0.279	0.223	……	0.192
本体	0.261	0.306	0.279	1.000	0.066	……	0.000
专利分析	0.203	0.257	2.223	0.066	1.000	……	0.000
……	……	……	……	……	……	……	……
模糊综合评价	0.189	0.214	0.192	0.104	0.000	……	1.000

4.4 综合共现相似度

得到直接共现相似度与间接共现相似度后，通过调整加权平均的权值进行多次实验，当权值均为 0.5 时实验效果最好，故对表 4 与表 5 中的相关性矩阵进行求和并取均值，得到研究方法类关键词综合共现相似度，如表 6 所示：

表 6 研究方法类关键词综合共现相似度

	知识图谱	可视化	社会网络	本体	专利分析	……	模糊综合评价
知识图谱	1.000	0.453	0.471	0.377	0.223	……	0.094
可视化	0.453	1.000	0.532	0.430	0.417	……	0.107
社会网络	0.471	0.532	1.000	0.318	0.404	……	0.096
本体	0.377	0.430	0.318	1.000	0.293	……	0.052
专利分析	0.223	0.417	0.404	0.293	1.000	……	0.000
……	……	……	……	……	……	……	……
模糊综合评价	0.094	0.107	0.096	0.052	0.000	……	1.000

4.5 层次结构建立

按照 3.2.3 小节所述步骤进行研究方法类关键词层次结构的建立。根据研究方法类关键词与特征项关键词的共现矩阵，若关键词与特征项共现次数在 1 及以上则认为其具有相关性。由此，统计与研究方法类关键词有关的特征项关键词个数，以表示该研究方法类关键词的概念范围，结果如图 2 所示：

ChinaXiv合作期刊
知识管理论坛，2022（4）:443-451
DOI: 10.13266/j.issn.2095-5472.2022.037



图 2 研究方法类关键词概念阈值分布

依据图 2 所示结果，“知识图谱”具有较大的概念范围，因此笔者选择“知识图谱”作为根节点进行研究方法类关键词层次结构的构建。此外，考虑到层级中关键词数量以及关键词概念范围的分布情况，笔者构建了具有 4 层层次关系的层次结构。对图 2 中关键词的概念范围分布情况进行分析，发现概念阈值在 22、15、8 等值附近波动较为明显，同时考虑到每一层级中的关键词节点数，设置第一层级的概念范围阈值为 22，第二层级的概念范围阈值为 15，第三层级的概念范围阈值为 8，第四层级的概念范围阈值为 1。

在加入子节点时基于对表 6 中相似度结果的分析，设置与根节点“知识图谱”综合共现相似度达到 0.15，即与根节点具有一定相关性的关键词能加入层次结构；基于对表 4 与表 5 中相似度结果的分析，设置与父节点间直接相似度或间接相似度达到 0.5，即与父节点具有较强相关性的关键词作为其子节点加入层次结构。基于表 6 中的结果，可以发现在 39 个研究方法类关键词中与根节点“知识图谱”综合共现相似度达到 0.15，可以加入层次结构的关键词共有 24 个。基于此，从根节点“知识图谱”开始依次向层次结构中加入子节点，根节点“知识图谱”作为层次结构的第一层级共有 3 个子节点，第二层级的 3 个节点共有 6 个子节点，第三层级的 6 个节点共有 5 个子节点，即可以加入层次结构的 24 个关键词中共有 14 个关键词加入层次结构，另有 10 个关键词与所有父节点均不满足相似度条件，故未加入层次结构。最后构建的以“知识图谱”为根节点的层次结构如图 3 所示：

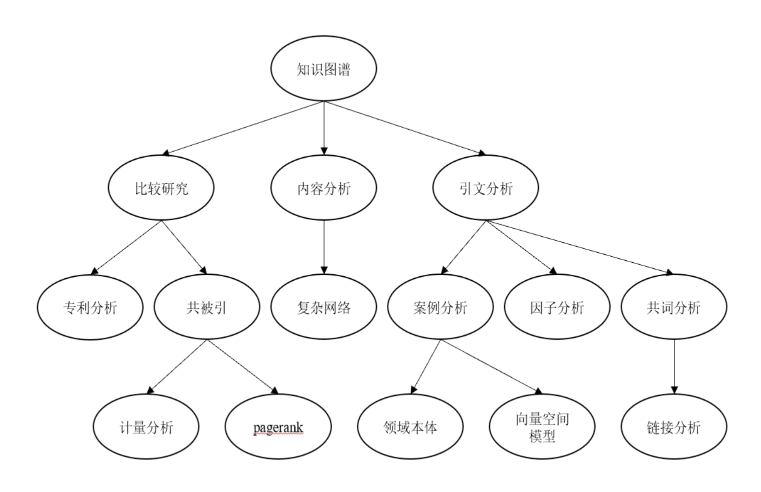


图 3 “知识图谱”层次结构

4.6 层次结构构建结果分析

为了与笔者提出的关键词层次结构构建方法进行对比，以“知识图谱”为根节点，分别基于方法类关键词间直接共现相似度和间接共现相似度构建层次结构。结果如图 4 所示：

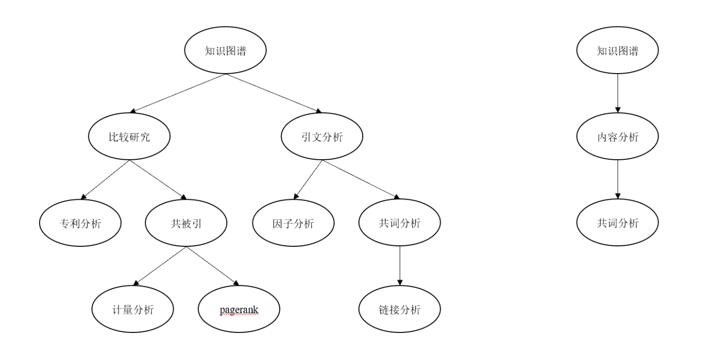


图 4 基于直接共现相似度和间接共现相似度的层次结构

比较图 3 和图 4 可以看出，基于单一共现指标的构建效果并不太理想，基于综合共现相似度的层次结构更为丰富，子节点较多，有利于建立完善的关键词层次结构。同时，图 3 所构建的层次结构将研究范围相似度较高的关键词联系起来，并且与同一研究主题相关度较高的关键词也聚集到一起，各关键词被分入到了合适的等级结构中。

5 结语

笔者以研究方法类关键词为研究对象，综合考虑关键词直接共现关系和间接共现关系，在关键词共现关系挖掘的基础上，分析与关键词关联的研究范围大小，建立了关键词层次结构。通过实例数据证明，笔者所提出的方法相较基于单一共现指标的方法，能够构建更为完善、关联更为紧密的关键词等级结构。但是，本文仍具有以下局限性：①关键词间间接共现存在多种情况，而本文仅考虑了两个研究方法类关键词应用于同一研究主题或研究范围的情况，未来将进一步探索多种间接关系的特点及其对关键词层次结构构建的影响；②受限于数据量，本文仅选用具有代表性的实例进行论证，如果选择的样本数据量较大，则更能充分体现关键词间的相互关系，那么层次结构构建的效果可能会更好。未来，笔者将在较大数据集合中对此层次结构构建方法予以验证。

参考文献：

[1] PUTRA J W G, KHODRA M L. Automatic title generation in scientific articles for authorship assistance: a summarization approach[J]. Journal of ICT research and applications, 2017, 11(3): 253-267.

[2] 罗威 , 谭玉珊 . 基于内容的科技文献大数据挖掘与应用 [J]. 情报理论与实践 , 2021, 44(6): 154-157.

[3] 胡昌平 , 陈果 . 科技论文关键词特征及其对共词分析的影响 [J]. 情报学报 , 2014, 33(1): 23-32.

[4] 韩普 , 王东波 , 朱恒民 . 基于复杂网络的汉语相似词挖掘和相似度计算研究 [J]. 情报学报 , 2015, 34(8): 885-896.

[5] 韩普 , 王东波 , 王子敏 . 词汇相似度计算和相似词挖掘研究进展 [J]. 情报科学 , 2016, 34(9): 161-165.

[6] 魏瑞斌 , 蒋倩雯 , 张瑞丽 . 基于文献共被引和共词分析的研究方法的比较研究——以共词分析和内容分析为例 [J]. 情报杂志 , 2019, 38(2): 36-42, 4.

[7] VARELAS G, VOUTSAKIS E, RAFTOPOULOU P, et al. Semantic similarity methods in wordnet and their application to information retrieval on the web[C]// Proceedings of the 7th annual ACM international workshop on Web information and data management. New York: ACM, 2005: 10-16.

[8] 田久乐 , 赵蔚 . 基于同义词词林的词语相似度计算方法 [J]. 吉林大学学报 (信息科学版), 2010, 28(6): 602-

DOI: 10.13266/j.issn.2095-5472.2022.037

608.
[9] 王义，王小林．基于改进的义原关联度算法的词语相关度计算 [J]. 情报学报，2012, 31(12): 1271-1275.
[10] 朱新华，马润聪，孙柳，等．基于知网与词林的词语语义相似度计算 [J]. 中文信息学报，2016, 30(4): 29-36.
[11] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL].[2022-07-31].https: //doi.org/10.48550/arXiv.1301.3781.
[12] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB/OL].[2022-07-31].https: //doi.org/10.48550/arXiv.1802.05365.
[13] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL].[2022-07-31].https: //doi.org/10.48550/arXiv.1810.04805.
[14] 田星，郑瑾，张祖平．基于词向量的 Jaccard 相似度算法 [J]. 计算机科学，2018, 45(7): 186-189.
[15] PONTES E L, HUET S, LINHARES A C, et al. Predicting the semantic textual similarity with siamese CNN and LSTM[EB/OL].[2022-07-31].https: //doi.org/10.48550/arXiv.1810.10641.
[16] SANJEEV M M, RAMALINGAM B, TK S K. Realtime semantic similarity analysis of bulk outlook Emails using BERT[C]//2020 International Conference on advances in computing, communication & materials (ICACCM). Piscataway: IEEE, 2020: 89-94.
[17] 闫强，张笑妍，周思敏．基于义原相似度的关键词抽取方法 [J]. 数据分析与知识发现，2021, 5(4): 80-89.
[18] TIBELY G, POLLNER P, VICSEK T, et al. Extracting tag hierarchies[J]. PloS one, 2013, 8(12): e84133.
[19] LI S, SUN Y, SOERGEL D. A new method for automatically constructing domain-oriented term taxonomy based on weighted word co-occurrence analysis[J]. Scientometrics, 2015, 103(3): 1023-1042.
[20] 熊回香，王学东．大众分类体系中标签概念空间的构建研究 [J]. 情报学报，2012, 31(9): 984-992.
[21] 熊回香，叶佳鑫．基于同义词词林的社会化标签等级结构构建研究 [J]. 情报杂志，2018, 37(1): 126-131.
[22] 叶佳鑫，熊回香，杨滋荣，等．关键词词频及语义特征对科技文献聚类的影响研究 [J]. 情报科学，2021, 39(8): 156-163.
[23] 孙鸿飞，侯伟，周兰萍，等．近五年我国情报学研究方法应用的统计分析 [J]. 情报科学，2014, 32(4): 77-84.
[24] 学术点滴，文献计量．COOC 一款用于文献计量和知识图谱绘制的新软件 [EB/OL]. [2021-07-15].https: //mp.weixin.qq.com/s/8RoKPLN6b1M5_jCk1J8UVg.

作者贡献说明：

熊回香：论文指导；
陈子薇：数据收集、论文撰写与修改；
叶佳鑫：论文修改。

Research on the Construction of Keyword Hierarchy Relationship Based on Co-occurrence Relationship

Xiong Huixiang Chen Ziwei Ye Jiaxin
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/Significance] Keyword is the most widely used literature knowledge unit, and the in-depth mining of its semantic relationship can provide underlying support for knowledge association and resource recommendation. **[Method/Process]** Based on the relationship between the direct co-occurrence and indirect co-occurrence of keywords, the correlation between keywords was mined, and on this basis, the distribution of keywords was analyzed, and the hierarchical structure between keywords was constructed according to the size of the concept range of keywords. **[Result/Conclusion]** Taking “knowledge graph” as the root node, this paper demonstrates the steps of construction of keywords hierarchy. The research shows that the method is feasible and effective, and it can construct the hierarchical structure of keywords better.

Keywords: keywords of scientific technological literature   keyword hierarchy   keyword characteristics